Title
Cluster Randomized Trials

ISBN
9781351620963

Authors
Crespi-Chun, Catherine
Crespi, Catherine

Publication Date
2019-04-24

Peer reviewed

Design and Analysis of Cluster Randomized Trials

Catherine M. Crespi
May 2019

Contents

1

Cluster Randomized Trials

CONTENTS

1.1 Introduction

In most randomized trials, individuals are randomized to study conditions and then followed to compare their outcomes. In a cluster randomized trial (CRT), also called a group randomized trial, preexisting groups of individuals such as clinics, schools or communities are randomized to conditions, with all individuals in the same group receiving the same treatment. After a follow-up period, outcomes are typically measured at the individual level. Examples of interventions that have been studied using a cluster randomized design include an intervention to improve the patient care experience for cancer patients that randomized nurses [63], an educational intervention to promote serologic testing for hepatitis B that randomized churches [5] and a cervical cancer screening program in rural India that randomized villages [54].

When individuals are randomized to conditions and do not interact with each other, their outcomes are generally regarded as independent. When preexisting groups are randomized, the outcomes of individuals in the same group cannot be considered independent. Rather, members of the same group share some commonalities — they may be patients with the same health care provider, children attending the same school or residents of the same village — and also may interact during the treatment period, which will make the outcomes of individuals in the same group more similar than the outcomes of individuals from different groups. As explained in Section 1.3.1.1, this correlation of outcomes within groups makes cluster randomized trials less statistically efficient (that is, the intervention effect estimates have larger standard errors) than individually randomized trials in which clustering does not occur. As a result, cluster randomized trials require larger overall numbers of individuals to achieve the same level of statistical power. They also require the use of data analysis methods that account for clustering.

If cluster randomized trials are less efficient, why use them? Various considerations may motivate the selection of a cluster randomized design. The intervention may naturally be implemented at the group level, e.g., group therapy or education sessions or a clinic-wide change in procedures. It may be less costly or logistically easier to implement the intervention at the cluster level. Cluster randomization can also prevent "contamination," that is, exposure of the control group to the intervention. Contamination tends to reduce differences in outcomes between conditions, making it more difficult to detect a treatment effect.

A key characteristic of cluster randomized trials is that they have multilevel data structure, with individuals at the lower level and clusters at the higher level. Thus design and analysis of cluster randomized trials are nat-

urally handled using multilevel modeling. Multilevel models are extendable to accommodate other modeling features such as covariates and additional hierarchical structure such as repeated measures on individuals or additional levels. In this chapter we emphasize the multilevel modeling approach to cluster randomized trial design and analysis. For general treatments of multilevel model analysis, see [26, 57].

This chapter is organized into three sections: randomization, analysis, and sample size and power. R code to implement methods is provided. Due to space constraints, we confine our attention to CRTs with two-level designs and a continuous or dichotomous outcome. Resources for CRTs with time-to-event or count outcomes or with more complex design elements such as additional levels, stepped wedge or crossover designs are provided in Section 1.5.

1.2 Randomization

Randomization helps to ensure balance across conditions on known and unknown prognostic factors. Due to randomization, we expect that the only systematic difference between two study arms will be that one received the intervention and the other did not; hence a comparison of the two conditions produces an unbiased estimate of the treatment effect.

Compared to an individually randomized trial, the number of randomized units in a cluster randomized trial is often relatively small, and simple randomization without restrictions (i.e., a coin flip) can result in chance imbalance between arms on important baseline covariates. For example, if 10 clinics are randomized to two conditions purely by chance, we may end up with most of the larger clinics in one arm and the smaller clinics in the other. Additionally, the number of clusters allocated to each arm may end up unequal. We briefly discuss several strategies for avoiding these problems, including matching and stratification, constrained randomization and minimization. For a more thorough discussion, see [23, 28].

1.2.1 Matching and stratification

In a matched or stratified design, clusters are sorted into groups or "strata" based on one or more prognostic factors, then clusters within strata are randomized to conditions [15, 23, 43]. A matched-pair design is the special case of strata of size 2; clusters are paired and one cluster in each pair is assigned to each condition. Randomization within strata defined by one or more characteristics ensures balance between arms on these characteristics.

The Korean Health Study [5] provides an example of the use of stratified randomization in a cluster randomized trial. This study evaluated a church-based intervention to improve hepatitis B virus serological testing among Ko-

rean Americans in Los Angeles. Fifty-two Korean churches were stratified by size (small, medium, large) and location (Koreatown versus other) and randomized to intervention or control conditions within the six strata. This ensured balance between the intervention and control arms on size and location. Church location was considered potentially prognostic because of acculturation differences among participants attending churches inside versus outside Koreatown. Church size was considered prognostic because of the potential for competing activities and resource differences at larger churches.

1.2.2 Constrained randomization

Stratification or matching become difficult when there are many matching or stratification factors and a limited number of clusters. Constrained randomization, also called restricted randomization, is an alternative [23, 47]. Constrained randomization involves generating all possible allocations of clusters to conditions, identifying the allocations that satisfy some predetermined balance criteria, then randomly selecting one allocation from the constrained set. This ensures acceptable balance on the predetermined criteria.

Constrained randomization was used in the implementation study reported by Maxwell et al. [38]. This study evaluated two strategies for implementing an evidence-based intervention to promote colorectal cancer screening in Filipino American community organizations. Twenty-two community organizations were randomized to either a basic or enhanced implementation strategy. Constrained randomization was used to ensure balance as well as to avoid contamination across arms. The investigators enumerated all two-group equal allocations of the 22 organizations that balanced the arms as to faith-based versus non-faith-based organizations, organizations with prior experience with the screening program versus organizations with no prior exposure, and zip code-level mean income and education, and also kept three organizations that were in close geographic proximity in the same arm (to prevent contamination), and randomly selected one of these allocations. The two groups were then randomly assigned to the basic or enhanced implementation strategy using a coin flip.

1.2.3 Minimization

Constrained randomization requires that all participating clusters be recruited and have relevant covariate information available at the beginning of the study. When clusters are recruited and randomized sequentially and/or there are many factors to balance, an alternative is minimization [59]. In minimization, the first few units (individuals in individually randomized trials, clusters in cluster randomized trials) are randomly assigned to conditions and subsequent units are randomized to the arm that will minimize an imbalance measure that considers multiple covariates. Although minimization has not been widely used

in cluster randomized trials [29], its ability to balance many covariates makes it an attractive option.

Randomization in cluster randomized trials

Cluster randomized trials typically involve a relatively small number of clusters, and as a result, simple unrestricted randomization can result in chance imbalance between arms on important prognostic factors. Techniques such as matching, stratification, constrained randomization and minimization can be useful for promoting balance across study arms.

1.3 Analysis

In this section, we discuss conducting the outcome analysis for cluster randomized trials. As discussed in the introduction, we focus on a multilevel modeling approach. We begin with a multilevel model for two-level data with a continuous outcome variable. This model introduces important concepts, including the intraclass correlation coefficient and the design effect. We then discuss estimation and inference for the intervention effect for continuous outcomes and for dichotomous outcomes.

Throughout this section, we assume that we have balanced data, meaning a two-arm trial with equal numbers of clusters in each condition and each cluster having an equal number of members. This assumption simplifies the derivation of key results. In practice, CRTs often have unequal numbers of clusters in each condition and clusters with varying numbers of members. In general, this does not alter the basic approach to estimation and inference using a general or generalized linear mixed effects model. However, these factors can have an impact on statistical power and sample size requirements. For this reason, we defer discussion of these issues to Section 1.4.

1.3.1 Continuous outcomes

Assume that we have continuous, normally distributed outcomes on individuals who are nested within clusters. For example, we may have pain scores on patients nested within hospital wards or depressive symptom scores on individuals nested within therapists. We set up a model for a single population of clusters and study some properties of the model. Then we add a covariate to encode cluster condition and discuss estimation and inference for the intervention effect.

1.3.1.1 Model

The basic model for a single population of clusters assumes that each cluster has its own mean and individuals within the cluster have outcomes that vary around that mean. The model for the outcome of individual i in cluster j, denoted Y_{ij}, is

$$Y_{ij} = \mu_j + \epsilon_{ij}, \tag{1.1}$$

where μ_j is the mean for cluster j and ϵ_{ij} is the error term indicating the discrepancy between the individual's observed outcome Y_{ij} and the cluster mean outcome μ_j.

We further assume that our clusters are sampled from a population of clusters that has an overall mean, with the cluster means varying around it. The model for the mean of cluster j, μ_j, is

$$\mu_j = \gamma_0 + u_j \tag{1.2}$$

where γ_0 denotes the population mean, assumed to be fixed, and u_j is a random effect representing the discrepancy between cluster j's mean and the population mean. Substituting equation (1.2) into equation (1.1) gives the single equation model

$$Y_{ij} = \gamma_0 + u_j + \epsilon_{ij}. \tag{1.3}$$

The random effects are assumed to be normal, with $u_j \sim N(0, \sigma_u^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, and to be independent of each other.

Model (1.3) is a simple model for two-level normally distributed data. Inspection of the model reveals that the total variance of an observation Y_{ij}, not conditional on cluster, can be decomposed as the sum of two independent variance components, one at the cluster level and the other at the individual level, namely,

$$Var(Y_{ij}) = \sigma_y^2 = \sigma_u^2 + \sigma_\epsilon^2. \tag{1.4}$$

Note that σ_ϵ^2 is the conditional variance of observations given that they are from the same cluster, and we expect that this variance will be lower than the total variance, i.e., $\sigma_\epsilon^2 \leq \sigma_y^2$.

A useful quantity for characterizing the apportionment of the total variance of the outcome between the two levels of variation is the intraclass correlation coefficient (ICC). The ICC, commonly denoted ρ, is defined as

$$\rho = \frac{\sigma_u^2}{\sigma_y^2} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}. \tag{1.5}$$

The ICC quantifies the proportion of the total variance of the outcome that is attributable to clustering, or more precisely, to variance of the cluster-level means. Because σ_u^2 and σ_ϵ^2 are non-negative, we must have $0 \leq \rho \leq 1$.

It can be shown that, for model (1.3), ρ also equals the correlation between two different observations from the same cluster, $Corr(Y_{ij}, Y_{i'j}), i \neq i'$.

Furthermore, the covariance of two different observations from the same cluster, $Cov(Y_{ij}, Y_{i'j})$, is equal to $\rho\sigma_y^2$. By rearranging (1.5), we also have that $\sigma_u^2 = \rho\sigma_y^2$ and $\sigma_\epsilon^2 = (1-\rho)\sigma_y^2$.

The intraclass correlation coefficient

For two-level data following model (1.3), the *intraclass correlation coefficient* (ICC), often denoted ρ, quantifies the proportion of the total variance of the outcome that is due to variance between clusters (i.e., variance in cluster-level means). The ICC is also equal to the correlation between two observations within the same cluster.

For most cluster randomized trials, ρ is small, typically in the range of 0.001 to 0.05. The value of the ICC in any specific trial will depend on the outcome variable, the type of cluster and other context-specific factors. Reporting guidelines recommend that cluster randomized trials report the observed ICC [9]. Reviews compiling ICC values from various studies include [11, 21, 44].

Now we consider sample means of data from cluster randomized trials and their variance. This will lead to some fundamental quantities and principles.

Suppose that we have a set of n_2 clusters with n_1 individuals in each cluster. This is our balanced data assumption. We discuss unequal allocation to conditions and varying cluster sizes in Section 1.4. The sample mean for cluster j can be calculated as

$$\bar{Y}_{\cdot j} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{ij}. \tag{1.6}$$

What is the variance of the sample cluster mean? Using rules for the variance of the sum of correlated random variables, the variance can be found to be

$$Var(\bar{Y}_{\cdot j}) = \frac{1}{n_1}\left[Var(Y_{ij}) + (n_1-1)Cov(Y_{ij}, Y_{i'j})\right] = \frac{\sigma_y^2}{n_1}\left[1 + (n_1-1)\rho\right]. \tag{1.7}$$

Let $\bar{Y}_{\cdot\cdot} = \frac{1}{n_2}\sum_{j=1}^{n_2}\bar{Y}_{\cdot j} = \frac{1}{n_1 n_2}\sum_{j=1}^{n_2}\sum_{i=1}^{n_1} Y_{ij}$ be the overall sample mean across all observations. Because observations in different clusters are assumed to be independent, the variance of $\bar{Y}_{\cdot\cdot}$ is

$$Var(\bar{Y}_{\cdot\cdot}) = \frac{\sigma_y^2}{n_1 n_2}\left[1 + (n_1-1)\rho\right]. \tag{1.8}$$

If the $n_1 n_2$ observations had been independent, the variance of the sample mean would have been $\sigma_y^2/(n_1 n_2)$. The ratio of the variances is $1 + (n_1-1)\rho$, which is called the design effect for cluster randomized trials. We expect that the design effect will be greater than one when $\rho > 0$; hence another common term for the design effect is the variance inflation factor. The design effect will

reduce to 1 when we have clusters of size 1 or when $\rho = 0$, that is, independent observations. The loss of statistical efficiency in cluster randomized trials is due to the design effect, which leads to larger standard errors.

The design effect

The term *design effect* comes from the field of survey sampling; see, for example, Kish [34]. When we conduct a cluster randomized trial, we can be regarded as collecting data from a *cluster sample* of individuals in each condition, rather than a simple random sample of individuals. The design effect or Deff quantifies the increase in the variance of the sample mean resulting from using a cluster sampling design:

$$Deff = \frac{\text{Variance for cluster sampling}}{\text{Variance for simple random sampling}}. \tag{1.9}$$

The design effect for a cluster randomized trial following model (1.3) equals $1+(n_1-1)\rho$ and represents the multiplicative factor by which the variance of the sample mean is increased due to cluster sampling.

Although ρ is typically small, the design effect also depends on cluster size and can be quite large. For example, an ICC of 0.02 and a cluster size of 100 leads to a variance inflation factor of 2.98, i.e., almost a tripling of the variance of the sample mean compared to independent observations. This represents a substantial loss of statistical efficiency.

1.3.1.2 Estimation and inference

Suppose now that our n_2 clusters are randomized to two conditions, with $n_2/2$ clusters in each condition. To accommodate different population means in each condition, we modify the model for the mean of cluster j to be

$$\mu_j = \gamma_0 + \gamma_1 w_j + u_j$$

where w_j is coded as -0.5 for the control condition and 0.5 for the intervention condition. Thus γ_0 is the grand mean (mean of the two means) and γ_1 is the difference in means between the two conditions. Note that the treatment indicator w_j is subscripted only by j and not by i, since treatment is assigned at the cluster level. The single equation model for the outcome Y_{ij} is

$$Y_{ij} = \gamma_0 + \gamma_1 w_j + u_j + \epsilon_{ij}. \tag{1.10}$$

We have allowed clusters in different conditions to have different cluster means but this was done using a fixed effect; we have not altered the random effect terms in the model. The total variance of an observation is still $Var(Y_{ij}) = \sigma_u^2 + \sigma_\epsilon^2$ and we still express the ICC as $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\epsilon^2)$, although it is possible that the magnitude of the variance components differs between conditions; we discuss this in Section 1.4.1.4.

Interest usually focuses on estimating the treatment effect γ_1. An unbiased estimate of γ_1 can be obtained as the difference of treatment group means. Indexing condition by $k = 1, 2$ and the outcomes as Y_{ijk}, our estimate of the treatment effect is

$$\hat{\gamma}_1 = \bar{Y}_{\cdot\cdot 1} - \bar{Y}_{\cdot\cdot 2}$$

with variance

$$Var(\hat{\gamma}_1) = \frac{4\sigma_y^2}{n_1 n_2}\left[1 + (n_1 - 1)\rho\right] \tag{1.11}$$

(the factor of 4 arises because there are $n_2/2$ clusters per condition). Using (1.4) and (1.5), we can also write this variance as

$$Var(\hat{\gamma}_1) = \frac{4}{n_1 n_2}\left(\sigma_\epsilon^2 + n_1\sigma_u^2\right). \tag{1.12}$$

The test for a treatment effect is the test of the null hypothesis $H_0 : \gamma_1 = 0$ in model (1.10). This test can be conducted using the test statistic

$$\frac{\hat{\gamma}_1}{SE(\hat{\gamma}_1)} \tag{1.13}$$

where $SE(\hat{\gamma}_1) = \sqrt{Var(\hat{\gamma}_1)}$. Under the null hypothesis, test statistic (1.13) has approximately a t distribution. In general, for a two-level model, the number of degrees of freedom (df) associated with the regression parameter for a cluster-level covariate is $n_2 - q - 1$, where q equals the total number of cluster-level covariates. For testing γ_1 in model (1.10), the df for the t statistic are $n_2 - 2$. When the df are large, the standard normal distribution can be used.

The parameter γ_1 and its standard error, as well as γ_0 and its standard error, can be estimated by maximum likelihood or restricted maximum likelihood (REML). When the number of clusters is small ($n_2 \leq 50$), REML estimation is recommended for estimating fixed effect parameters; for datasets with a larger number of clusters, either method may be used and should give similar results [37, 57]. REML estimates are preferred for estimating the variance components σ_ϵ^2 and σ_u^2, regardless of dataset size, because maximum likelihood estimators of the variance components have a downward bias [57].

If stratification is used, the model should include indicators for strata, which ensures that the treatment effect estimates are conditional on stratum. Stratification can increase power; see Section 1.4.1.8.

1.3.1.3 Example

To illustrate inference for a cluster randomized trial with a continuous outcome, we simulate and analyze data based loosely on a CRT of a pain self-management intervention for cancer patients reported in Jahn et al. [30]. The intervention was delivered in the hospital setting and involved nurse-led counseling program focused principally on reducing patient-related cognitive barriers. To avoid contamination across conditions, the study was designed as a

cluster randomized trial and the intervention was applied at the ward level. Nurses in wards assigned to the intervention condition received special training, while nurses in control wards did not. Outcomes were measured on patients in the wards. The primary outcome was patient score on the Barriers Questionnaire II.

While the actual trial had 9 oncology wards in each condition and a variable number of patients per ward, for pedagogical reasons we simulated data with 10 wards per condition and 10 patients per ward. Our simulated data were based on model (1.10) with parameter values $\gamma_0 = 60$, $\gamma_1 = 10$, $\sigma_u^2 = 25$ and $\sigma_\epsilon^2 = 600$. These values imply that $\sigma_y^2 = 625$ and $\rho = 0.04$. The R commands to simulate the data and fit the model are

```
# set parameter values
n2 <- 20
n1 <- 10
gamma_0 <- 60
gamma_1 <- 10
sigma_u <- sqrt(25)
sigma_e <- sqrt(600)
sigma_y <- sqrt(625)

# simulate data
set.seed(96135)
u <- rep(rnorm(n2, sd=sigma_u), each=n1)
e <- rnorm(n1*n2, sd=sigma_e)
w <- c(rep(-0.5, n1*n2/2), rep(0.5, n1*n2/2))
y <- gamma_0 + gamma_1*w + u + e
j <- rep(seq(1:n2), each=n1)
pain.data <- data.frame(y, j, w)

# fit model
library(lme4)
paincrt <- lmer(y ~ w + (1|j), data=pain.data)
summary(paincrt)
...
Random effects:
 Groups   Name        Variance Std.Dev.
 j        (Intercept)  30.3     5.51
 Residual             618.7    24.87
Number of obs: 200, groups:  j, 20

Fixed effects:
            Estimate Std. Error    df t value Pr(>|t|)
(Intercept)    63.95       2.15 18.00   29.79   <2e-16 ***
w               9.90       4.29 18.00    2.31    0.033 *
```

The grand mean is estimated as $\hat{\gamma}_0 = 63.95$ while the treatment effect is estimated as $\hat{\gamma}_1 = 9.90$ and is significant at the 0.05 level. The df for both fixed effects parameters is $20 - 2 = 18$. The estimated variance components are $\hat{\sigma}_u^2 = 30.3$ and $\hat{\sigma}_\epsilon^2 = 618.7$, from which we can calculate $\hat{\sigma}_y^2 = 649.0$ and $\hat{\rho} = 0.047$. The estimated parameter values do not coincide with the true values due to sampling variability.

What would we have inferred if we had neglected to account for the clustering of observations within wards? The following R code fits the linear regression model $Y_{ij} = \gamma_0 + \gamma_1 w_j + e_{ij}$, which assumes independent observations, to the data:

```
# Wrong model!
summary(lm(y ~ w , data=pain.data))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     64.0        1.8   35.58   <2e-16 ***
w                9.9        3.6    2.75   0.0064 **
```

We get the same point estimates for γ_0 and γ_1, but the standard errors are substantially smaller and hence the p-value for the treatment effect is also lower. Although in this case we would have reached the same conclusion about rejecting the null hypothesis of no treatment effect at level 0.05, in many data analyses, the deflated p-value would have led us to incorrectly reject the null. The regression parameter estimates are identical in the models fit with and without clustering because our data are balanced; in the case of unbalanced data, this will not always occur.

We can get confidence intervals for the variance components using the command

```
confint(paincrt)
```

which yields a 95% confidence interval for $\hat{\sigma}_u$ of $(0.00, 10.2)$. The interval appears to include zero. Should we drop this variance component? No, as explained in the accompanying box.

Should nonsignificant variance components be dropped?

When analyzing data from a cluster randomized trial, if a variance component is not significantly different from zero, should it be dropped from the model?

Even a small ICC, if ignored, can inflate the Type I error rate, that is, the probability that we erroneously reject the null hypothesis and declare the intervention to be effective. Furthermore, the standard errors for variance components are not well estimated when their true values are close to zero, and the degrees of freedom for such tests, which are based on the number of clusters, are usually limited, which limits the power of such tests. Therefore it is recommended that all random effects associated with the study design and sampling plan be retained in the model.

In the example, the true value of the variance of the cluster means is known to be non-zero because the data were simulated. Dropping this term would lead to a misspecified model.

1.3.2 Dichotomous outcomes

Now we consider cluster randomized trials with dichotomous outcomes. Examples of dichotomous outcomes include achieving a tumor response, receiving a cancer screening procedure or acquiring an infection.

There are two common approaches to modeling dichotomous outcomes from cluster randomized trials [16]. One approach models the cluster-level proportions, and the second models the cluster-level log odds. We discuss both approaches. We spend some time discussing the intraclass correlation, which is more complicated for dichotomous data than it is for continuous data. We first discuss simple models for two-level clustered data, without covariates or different intervention conditions, in order to study important principles, and then discuss estimation and inference for an intervention effect.

1.3.2.1 Cluster-level proportions model

Let Y_{ij} denote the dichotomous outcome of the ith individual in the jth cluster, where $Y_{ij} = 1$ for success and 0 for failure. Under the cluster-level proportions model, the individuals in cluster j have a probability of success that is specific to their cluster, denoted π_j. Thus the Y_{ij} are Bernoulli random variables with success probability π_j. The cluster-level success probabilities π_j are assumed to be random variables that follow a distribution with $E(\pi_j) = \pi$ and $Var(\pi_j) = \sigma_d^2$. The specific distribution does not affect the key results and we leave it unspecified. Under this model, the mean and variance of Y_{ij}, unconditional on cluster, are $E(Y_{ij}) = E(\pi_j) = \pi$ and $Var(Y_{ij}) = \pi(1 - \pi)$. This leads to an expression for the ICC in the cluster-level proportions model

as

$$\rho_d = \frac{Var(\pi_j)}{Var(Y_{ij})} = \frac{\sigma_d^2}{\pi(1-\pi)}. \tag{1.14}$$

Now we consider sample proportions and their properties. The sample proportion for cluster j can be calculated as

$$\hat{\pi}_j = \bar{Y}_{\cdot j} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{ij},$$

where n_1 is cluster size. The sample cluster proportion is an unbiased estimate of the true cluster proportion, $E(\hat{\pi}_j) = \pi_j$, and the variance of $\hat{\pi}_j$ can be found to be

$$Var(\hat{\pi}_j) = \frac{\pi(1-\pi)}{n_1}\left[1 + (n_1 - 1)\rho_d\right].$$

This variance is the analogue of the variance of the sample cluster mean for continuous outcomes given in equation (1.7).

Let $\hat{\pi} = \bar{Y}_{\cdot\cdot} = \frac{1}{n_1 n_2}\sum_{j=1}^{n_2}\sum_{i=1}^{n_1} Y_{ij}$ be the overall sample proportion across all observations, assuming n_2 clusters of size n_1. The overall sample proportion provides an unbiased estimate of population proportion; $E(\hat{\pi}) = \pi$. Because observations in different clusters are assumed to be independent, the variance of $\hat{\pi}$ is

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n_1 n_2}\left[1 + (n_1 - 1)\rho_d\right]. \tag{1.15}$$

Had the $n_1 n_2$ observations been independent, the variance of the sample proportion would have been $\frac{\pi(1-\pi)}{n_1 n_2}$. As for continuous outcomes, the ratio of the variances is the design effect, $1 + (n_1 - 1)\rho_d$, which reflects the increase in the variance of the sample proportion attributable to correlation of observations within clusters, or from another perspective, due to cluster sampling of observations.

Now suppose that our n_2 clusters are randomized to two conditions with $n_2/2$ clusters in each condition. Denote the success proportions in the two conditions as π_1 and π_2. Adding a subscript k to denote condition, we could estimate these proportions as $\hat{\pi}_k = \bar{Y}_{\cdot\cdot k} = \frac{1}{n_1 n_2/2}\sum_{j=1}^{n_2/2}\sum_{i=1}^{n_1} Y_{ijk}$ for $k = 1, 2$. The intervention effect can be estimated as the difference in sample proportions, $\hat{\pi}_1 - \hat{\pi}_2$, and its variance is

$$Var(\hat{\pi}_1 - \hat{\pi}_2) = \left[\frac{\pi_1(1-\pi_1)}{n_1 n_2/2} + \frac{\pi_2(1-\pi_2)}{n_1 n_2/2}\right]\left[1 + (n_1 - 1)\rho_d\right]. \tag{1.16}$$

This result is the basis for a commonly used sample size calculation approach for cluster randomized trials with dichotomous outcomes, which we discuss in section 1.4.2.1. However, the cluster-level log odds model, which we discuss next, is more commonly used for analysis.

1.3.2.2 Cluster-level log odds model

The other approach for modeling dichotomous outcomes for CRTs is to use a random effects logistic regression model, such as the logistic-normal model. This model assumes that Y_{ij} is Bernoulli with cluster-specific success probability π_j, and that the logits of cluster proportions π_j follow a normal distribution. The basic model without covariates can be expressed as

$$log\left(\frac{\pi_j}{1-\pi_j}\right) = \gamma_0 + u_j \tag{1.17}$$

where $u_j \sim N(0, \sigma_u^2)$.

Under this model, the between-cluster variance σ_u^2 is expressed on the log-odds scale. The model implicitly assumes an overall population proportion $\pi = 1/(1 + e^{-\gamma_0})$, making the total population outcome variance equal to $\pi(1-\pi)$, which is on the proportions scale. The quantities σ_u^2 and $\pi(1-\pi)$ are not comparable because they are on different scales and it is not sensible to form an ICC as their ratio. One way of finding an ICC for model (1.17) that is on the proportions scale is to use a Taylor expansion of $\text{logit}(\pi_j)$, which yields an approximation of ρ_d as

$$\rho_d \approx \sigma_u^2[\pi(1-\pi)]^2; \tag{1.18}$$

see [61]. An alternative approach is to define the ICC on the log-odds scale. This leads to the expression

$$\rho_{d(l)} = \frac{\sigma_u^2}{\sigma_u^2 + \Pi^2/3} \tag{1.19}$$

where Π is the mathematical constant 3.14156...; see [16, 57]. The term $\Pi^2/3$ is the variance of the standard logistic distribution and plays the role of the within-cluster variance.

ICCs for dichotomous data

The ICC for clustered dichotomous data can be defined on the proportions scale (ρ_d) or the log odds scale ($\rho_{d(l)}$). These ICCs can take very different values for the same data. The proportional discrepancy between ρ_d and $\rho_{d(l)}$ is greater for larger values of ρ_d and when the prevalence π is farther from 0.5. For further discussion, see [16].

Researchers should be aware of the two different scales for the ICC for clustered dichotomous data and be careful to use the ICC on the correct scale for sample size and power calculations.

1.3.2.3 Estimation and inference

To model data from a cluster randomized trial, we use the cluster-level log odds model and expand the model to include a covariate encoding cluster condition. The model is

$$log\left(\frac{\pi_j}{1-\pi_j}\right) = \gamma_0 + \gamma_1 w_j + u_j \tag{1.20}$$

where w_j is coded as -0.5 for the control condition and 0.5 for the intervention condition. Thus γ_0 is the average log odds of success across all clusters and γ_1 is the difference in log odds for success between the two conditions. The intervention effect is typically reported as an odds ratio, obtained as e^{γ_1}.

A closed form expression for the estimator $\hat{\gamma}_1$ and its variance can be derived; see [42, 58]. Assuming equal-sized clusters of size n_1 and $n_2/2$ clusters per condition, the intervention effect can be estimated as the difference in average log odds between conditions and the variance of $\hat{\gamma}_1$ can be estimated as

$$Var(\hat{\gamma}_1) = \frac{4(\sigma_u^2 + \tau^2/n_1)}{n_2} \tag{1.21}$$

where

$$\tau^2 = \frac{1}{2}\left[\frac{1}{\pi_1(1-\pi_1)} + \frac{1}{\pi_2(1-\pi_2)}\right] \tag{1.22}$$

is a measure of variability at the individual level.

There are various algorithms for fitting mixed-effects logistic models and obtaining parameter estimates, standard errors and confidence intervals [33]. The expression for the likelihood of a mixed-effects model is an integral over the random effects space. For a linear mixed-effects model, this integral can be evaluated exactly. For a generalized linear mixed-effect models, the integral must be approximated. Different approximation methods can give slightly different results.

1.3.2.4 Example

To illustrate analysis for a cluster randomized trial with a dichotomous outcome, we simulate data based on the Breast Cancer Education Program for Samoan Women [39]. This study evaluated the effectiveness of a breast cancer education program tailored to women with Samoan ancestry in the United States. In the trial, 61 Samoan churches were randomized to the intervention or control condition. Women from churches in the intervention arm participated in culturally tailored interactive group discussion sessions with a health educator; the control condition was usual care. The primary outcome was self-reported receipt of a mammogram within eight months.

The trial had a variable number of women per church, ranging from 1 to 42 with a median of 13. For simplicity, we simulated data with 30 churches in each condition and 12 participants per church. The trial had a rather high ICC

of 0.19 on the proportions scale (ρ_d). We simulated data with $\rho_d = 0.1$. We assume that the proportions of participants with self-reported mammogram receipt are 0.30 and 0.50 in the control and intervention arms, respectively. The overall population proportion is thus 0.40, and using equation (1.18), we find that we need to set $\sigma_u = 1.318$. Using equation (1.19), the ICC on the log odds scale is $\rho_{d(l)} = 0.345$ (note that this value differs substantially from the ρ_d of 0.1). The regression coefficients needed to reflect the success proportions in each arm are $\gamma_0 = -0.4237$ and $\gamma_1 = 0.8473$. R code to simulate data is

```
# set parameter values
n2 <- 60
n1 <- 12
gamma_0 <- -0.4236
gamma_1 <- 0.8473
sigma_u <- 1.318

# simulate outcome data
set.seed(32410)
linprobs1 <- gamma_0-gamma_1/2+rep(rnorm(n2/2, sd=sigma_u), each=n1)
linprobs2 <- gamma_0+gamma_1/2+rep(rnorm(n2/2, sd=sigma_u), each=n1)
linprobs <- c(linprobs1, linprobs2)
probs <- 1/(1+exp(-linprobs))
y <- sapply(probs, function(x) sample(0:1, 1, prob = c(1-x, x)))
j <- rep(seq(1:n2), each=n1)
w <- c(rep(-0.5, n1*n2/2), rep(0.5, n1*n2/2))
mamm.data <- data.frame(y, j, w)

# average success proportions in each condition
mean(mamm.data$y[mamm.data$w==0.5])
mean(mamm.data$y[mamm.data$w==-0.5])
```

The success proportions in the simulated data are 0.36 and 0.49 in the control and intervention arms, respectively. Figure 1.1 shows the distribution of the cluster proportions and the logits of the cluster proportions in the two study arms. The proportions and logits are shifted somewhat lower for the control clusters. The logits of the cluster proportions are specified as normally distributed with different means in each condition; the normality assumption is not fully apparent in the figure, particularly in the control arm, due to the relatively low number of clusters in each condition.

R code to fit the model using glmer in R and abbreviated output are below:

```
# fit model
summary(glmer(y ~ w + (1 | j), data = mamm.data, family = binomial))
...
Random effects:
 Groups Name        Variance Std.Dev.
```

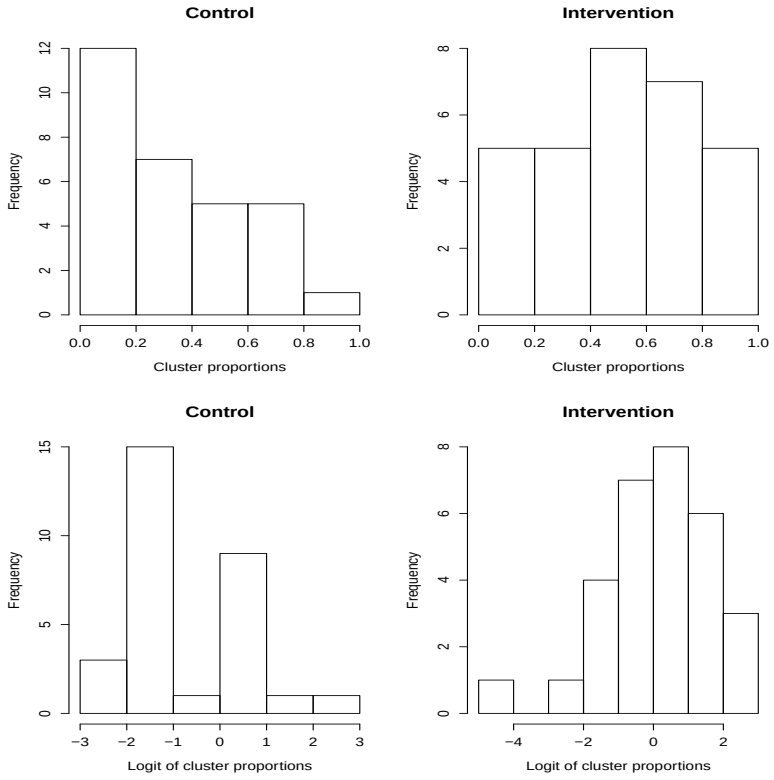

FIGURE 1.1
Distribution of cluster-level proportions and logits of the cluster-level proportions from simulated data, by intervention condition.

```
 j      (Intercept) 1.71     1.31
Number of obs: 720, groups:  j, 60

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.418      0.192   -2.18    0.029 *
w              0.694      0.384    1.81    0.071 .
```

The test statistics for the fixed effects have approximately a normal distribution rather than a t distribution under the null, so there are no df to consider. The estimates differ from the true values due to sampling variability. The p-value exceeds the benchmark of 0.05. The odds ratio for the treatment effect is $e^{0.694} = 2.0$. The estimated standard deviation of the random effect is 1.31. Using (1.19) and (1.18), we can calculate the estimated ICCs as 0.342 on the log odds scale and 0.098 on the proportions scale. These are close to the true

values. The discrepancy between these two ICCs serves as a reminder that these two quantities are on different scales and should not be confused.

1.3.3 Other analysis methods

We have discussed analysis of data from CRTs using multilevel modeling (general or generalized linear mixed models). This approach is statistically efficient and easily accommodates regression adjustment for covariates or specification of additional hierarchical data structure. Other methods may be useful for specific studies. For example, a two-sample t test comparing cluster-level summary statistics (means, proportions) is robust to departures from the normality assumption [23]. Other robust options include nonparametric tests on cluster-level statistics and permutation tests. Some of these methods allow for a limited amount of covariate adjustment. For further information, see [23].

Another approach is generalized estimating equations (GEE) [14, 35]. GEE assumes a linear or generalized linear model for the expected values of the dependent variable, conditional on the explanatory variables, but does not fully specify a probability model for the data. Rather, the parameters are estimated under a "working model" for the covariance structure; for a cluster randomized trial, an exchangeable correlation structure is typically assumed. Standard errors are obtained using a robust sandwich estimator. For our example, the R code and abbreviated output are

```
summary(geeglm(y ~ w, id=j, data=mamm.data, family=binomial,
   corstr= "exchangeable"))
...
 Coefficients:
            Estimate Std.err Wald Pr(>|W|)
(Intercept)   -0.319   0.148 4.64    0.031 *
w              0.527   0.296 3.17    0.075 .
...
Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.253  0.0476
```

The coefficient estimates in a GEE model are population-average estimates; here the estimated population-average odds ratio is $e^{0.527} = 1.7$. In contrast, mixed-effects logistic models provide odds ratios conditional on cluster. In general, population-averaged odds ratios are closer to the null than are cluster-conditional odds ratios. However, the p-values tend to be similar. Here, the p-values are close. The GEE estimated correlation parameter is the Pearson correlation between observations in the same cluster. For further discussion of population-average versus cluster-specific approaches, see [19, 27].

1.4 Sample Size and Power

When designing a cluster randomized trial or other study, we typically want to ensure that the sample size will be adequate to achieve the study's objectives. The primary objective is usually to detect a clinically meaningful and statistically significant difference between outcomes in the intervention and control conditions.

This section discusses how to calculate statistical power for a cluster randomized trial with a continuous or binary outcome, and how to find the sample size required to achieve a desired level of power. For both types of outcomes, results are first derived for the case of balanced data, with equal numbers of clusters in each condition and clusters of equal size. Subsequent sections consider unequal allocation of clusters, varying cluster sizes, and unequal ICCs in the two arms.

We restrict attention to cluster randomized trials with two-level data structure and a continuous or binary outcome. Discussion of power and sample size for other cluster randomized trial designs and other types of outcomes can be found in [8, 12, 15, 23, 41, 43, 53].

1.4.1 Continuous outcomes

1.4.1.1 Power

We begin by assuming that our data follow the two-level normal model (1.10), under which the observation Y_{ij}, for individual i in cluster j, follows

$$Y_{ij} = \gamma_0 + \gamma_1 w_j + u_j + \epsilon_{ij}$$

with $u_j \sim N(0, \sigma_u^2)$, $e_{ij} \sim N(0, \sigma_\epsilon^2)$, and u_j and ϵ_{ij} independent. We test for an intervention effect by testing $H_0 : \gamma_1 = 0$ using test statistic (1.13). When the null hypothesis is true, the test statistic follows a t distribution with $n_2 - 2$ degrees of freedom. When $\gamma_1 \neq 0$, the test statistic follows a noncentral t distribution (see box), which has two parameters, a df and a noncentrality parameter. Here, the df are $n_2 - 2$ and when there are $n_2/2$ clusters in each condition, n_1 individuals in each cluster and equal ICCs in each arm, the noncentrality parameter is

$$\lambda = \frac{\gamma_1}{\sqrt{\frac{4(\sigma_\epsilon^2 + n_1\sigma_u^2)}{n_1 n_2}}} = \frac{\gamma_1}{\sqrt{\frac{4\sigma_y^2[1+(n_1-1)\rho]}{n_1 n_2}}}. \tag{1.23}$$

The numerator is the true difference in means between conditions and the two versions of the denominator are the square root of the variance of $\hat{\gamma}_1$; see Section 1.3.1.2.

Noncentral t distribution

A random variable of the form

$$\frac{Z+\lambda}{\sqrt{\frac{\chi^2_\nu}{\nu}}} \tag{1.24}$$

where Z is a standard normal random variable, χ^2_ν is a chi-square random variable with ν degrees of freedom and λ is a constant, has a noncentral t distribution with ν degrees of freedom and noncentrality parameter λ, denoted $t_{\nu,\lambda}$. The standard (central) t distribution is the special case of $\lambda = 0$.

It is often convenient to work with standardized effect sizes, which give the difference between the intervention and control condition means in units of the standard deviation of the outcome variable. Here, the standardized effect size is $\delta = \gamma_1/\sigma_y$, where σ^2_y is the total variance of the outcome. The noncentrality parameter can then be expressed as

$$\lambda = \frac{\delta}{\sqrt{\frac{4[1+(n_1-1)\rho]}{n_1 n_2}}}. \tag{1.25}$$

Benchmarks for standardized effect sizes are given by Cohen [10], who suggested that 0.2, 0.5 and 0.8 represent small, moderate and large effect sizes.

The power $1-\beta$ of a hypothesis test is the probability that the value of the test statistic is more extreme than the critical value(s) given that some specified scenario is true. For a two-sided test with Type I error rate α, the power for a CRT following the two-level normal model (1.10) can be calculated as

$$P[t_{n_2-2,\lambda} > t_{n_2-2,0}(1-\alpha/2)] + P[t_{n_2-2,\lambda} < t_{n_2-2,0}(\alpha/2)] \tag{1.26}$$

where $t_{\nu,0}(a)$ denotes the ath quantile of the standard t distribution. One of these tail probabilities will typically be very small and can be neglected. When the noncentrality parameter λ is expressed as in equation (1.23), the parameter values required to compute power are n_1, n_2, γ_1, and either σ^2_ϵ and σ^2_u or σ^2_y and ρ. When λ is expressed as in (1.25), the parameter values required are n_1, n_2, δ and ρ.

An R function to calculate power for a two-level, normal outcome CRT with balanced data and a two-sided test is

```
power.crt.bal <- function(delta, rho, n1, n2, alpha){
  ta <- qt(1-alpha/2, n2-2)
  tb <- qt(alpha/2, n2-2)
  deff <- 1 + (n1-1)*rho
  lambda <- delta/sqrt(4*deff/(n1*n2))
  df <- n2-2
```

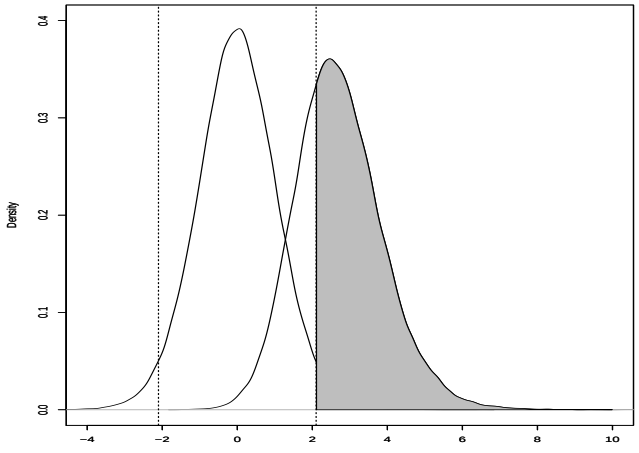

FIGURE 1.2
Illustration of power calculation. Parameter values are $\delta = 0.4$, $\rho = 0.02$, $n_1 = 10$ and $n_2 = 20$. Density on the left represents the distribution of the test statistic under the null, a (central) t distribution with 18 df. Density on the right represents its distribution under the alternative, a noncentral t with 18 df and noncentrality parameter 2.604. Dashed vertical lines indicate the critical values. Shaded area represents power.

```
  pow <- pt(ta, df, lambda, lower.tail=FALSE) + pt(tb, df, lambda)
  print("Deff is")
  print(deff)
  print("Power is")
  return(pow)
}
power.crt.bal(0.4, 0.02, 10, 20, 0.05)
```

Using this function, we can find that the power for a trial with an effect size of $\delta = 0.4$, $\rho = 0.02$, $n_1 = 10$ individuals per cluster and $n_2 = 20$ total clusters is 0.69 and the design effect is 1.18. The relationship between the distribution of the test statistic under the null and alternative hypotheses for this example is depicted in Figure 1.2.

What factors affect the power of a cluster randomized trial? As $|\lambda|$ increases, the noncentral t distribution moves farther away from zero and power increases. The factors affecting power can thus be gleaned from the expressions for the noncentrality parameter in (1.23) and (1.25). Power increases as the treatment effect $|\gamma_1|$ increases and decreases as any of the variance parameters σ_y^2, σ_u^2 or σ_ϵ^2 or ρ increase, all else being equal. What happens to power as we increase cluster size n_1 or number of clusters n_2? To investigate this,

we rewrite $Var(\hat{\gamma}_1)$ as

$$Var(\hat{\gamma}_1) = 4\left(\frac{\sigma_\epsilon^2}{n_1 n_2} + \frac{\sigma_u^2}{n_2}\right). \tag{1.27}$$

As n_2 increases, both components of the variance decrease; as $n_2 \to \infty$, $Var(\hat{\gamma}_1) \to 0$, $\lambda \to \pm\infty$ and power $\to 1$. However, increasing cluster size n_1 only reduces the first component; it has no effect on the influence of the variance of the cluster means. As $n_1 \to \infty$, $Var(\hat{\gamma}_1) \to \sigma_u^2/n_2$. Thus at some point, increasing the number of individuals per cluster will have a negligible effect on power. In general, power for CRTs is driven more by number of clusters than by cluster size.

Power: number of clusters versus cluster size

In general, the power of a cluster randomized trial is influenced more strongly by the number of clusters than by the number of individuals per cluster. To increase power, increasing the number of clusters is usually a more effective strategy than increasing cluster size.

1.4.1.2 Sample size: number of clusters

Suppose that we wish to determine the number of clusters required to achieve a desired level of power. The size of the clusters is assumed known and constant. Equation (1.26) provides power as a function of total number of clusters n_2; however, we cannot simply invert the equation and solve for n_2 as a function of power because n_2 appears in both the noncentrality parameter of the t distribution and the degrees of freedom. However, the equation can be solved iteratively until the minimum n_2 that provides sufficient power is identified. Note that for equal allocation of clusters to study arms, n_2 must be an even number.

When the number of clusters and therefore the df are sufficiently large ($n_2 \geq 30$ or so), the normal approximation to the t can be used. Using this approach, the minimum total number of clusters required, with $n_2/2$ in each condition, can be calculated as

$$n_2 \geq 4\frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{n_1\delta^2}\left[1 + (n_1 - 1)\rho\right], \tag{1.28}$$

where z_p represents the *pth* quantile of the standard normal distribution and we assume a two-sided test with Type I error rate α. The calculated value of n_2 will need to be rounded to the next highest even integer to achieve an equal number of clusters in each arm. The following R function computes the sample size:

```
sampsize.crt.bal <- function(delta, rho, n1, beta, alpha){
```

```
  za <- qnorm(1-alpha/2)
  zb <- qnorm(1-beta)
  n2 <- 4*(za+zb)^2*(1+(n1-1)*rho)/(n1*delta^2)
  print("Total number of clusters required calculated as")
  print(n2)
  print("Required clusters per arm is")
  print(ceiling(n2/2))
}
```

Example. Suppose we wish to find the minimum number of clusters required to detect a small effect size of 0.2 with 80% power assuming an ICC of 0.05, clusters of size 25 and Type I error rate of 0.05. Using the command

```
sampsize.crt.bal(0.2, 0.05, 25, 0.2, 0.05)
```

we find that n_2 is calculated as 69.1 clusters, which we round up to 70 total clusters (35 per condition) to ensure at least 80% power.

Some authors suggest that when the number of clusters and therefore the df for the t distribution are small, one additional cluster per arm should be added to the value calculated by equation (1.28) to account for using the normal rather than the t distribution; see [23]. In general, a more accurate calculation can be performed using the power equation (1.26) iteratively.

Example. To detect a larger effect size of $\delta = 0.6$, using equation (1.28), we calculate $n_2 \geq 7.7$. This number is low enough that we are concerned about using the normal approximation to the t distribution. Using power formula (1.26) to get a more precise result, we calculate that for 8 total clusters the power is 67%, whereas for 10 total clusters the power is 80%. Thus we need 10 clusters in total.

If we drop the design effect from sample size formula (1.28) and rearrange terms, we arrive at

$$N = n_1 n_2 \geq 4\frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}, \tag{1.29}$$

which provides the minimum required total sample size N for a two-sample t test with equal variances and equal-sized samples $n_1 n_2/2$ in each group, assuming large samples. The difference is the design effect. A common rubric for calculating the sample size required for a cluster randomized trial is to calculate the sample size for a trial involving independent observations and then inflate by the design effect. This approach leads to formula (1.28).

Variance inflation approach to sample size calculation for cluster randomized trials

A common approach for calculating the sample size for a cluster randomized trial to achieve a desired level of power is to calculate the sample size requirement assuming independent observations and then inflate this number by the design effect:

$$\begin{aligned}&\text{Total N for cluster randomized trial}\\ &\quad\approx \text{Total N for individually randomized trial} \times \text{Deff.}\end{aligned} \tag{1.30}$$

While this approach can give a good approximation, it is recommended that other factors that may affect power be considered, such as varying cluster sizes, unequal ICCs and covariates.

Example. Using the variance inflation approach, for $\delta = 0.2$ for an individually randomized trial, we need 785 total subjects to achieve 80% power. For $n_1 = 25$ and $\rho = 0.05$, the design effect is $1 + (n_1 - 1)\rho = 2.2$, so we inflate the total sample size to 1,727. Clusters are of size 25, so this total number of observations corresponds to 69 clusters, which we round up to a total of 70, or 35 in each arm.

1.4.1.3 Sample size per cluster

In some situations, we may have a fixed number of clusters available but have a choice as to the number of observations to sample from each cluster. To determine the number of observations to sample from each cluster to achieve a desired level of power, we can solve (1.28) for n_1, which yields

$$n_1 = \frac{4(1-\rho)}{\frac{n_2\delta^2}{(z_{1-\alpha/2}+z_{1-\beta})^2} - 4\rho}. \tag{1.31}$$

An R function to calculate sample size per cluster is

```
sampsizeper.crt.bal<-function(delta, rho, n2, beta, alpha){
  za <- qnorm(1-alpha/2)
  zb <- qnorm(1-beta)
  n1 <- 4*(1-rho)/((n2*delta^2/(za+zb)^2) - 4*rho)
  return(n1)
}
```

Example. Suppose we have 12 hospitals willing to participate in a study and we wish to know how many patients to sample from each hospital to achieve 80% power to detect an effect size of $\delta = 0.5$, assuming $\rho = 0.05$. The R command

```
sampsizeper.crt.bal(0.5, 0.05, 12, 0.2, 0.05)
```

calculates $n_1 = 20.9$, indicating that we need 21 patients per cluster.

It will not always be possible to achieve a desired level of power with a fixed number of clusters, even when using an arbitrarily large cluster size. Indeed, the solution for n_1 will be negative if $n_2 < 4\rho(z_{1-\alpha/2} + z_{1-\beta})^2/\delta^2$, reflecting the impossibility of always achieving desired power by increasing cluster size. The influence of cluster number versus cluster size on power is discussed in Section 1.4.1.1.

1.4.1.4 Unequal ICCs in treatment arms

Thus far we have made the implicit assumption that σ_y^2 and ρ, or equivalently σ_u^2 and σ_ϵ^2, are the same across arms. In some studies, we might expect the variance or correlation parameters to differ across arms. For example, an intervention that encourages interaction among cluster members may result in a higher ICC in the intervention arm, or heterogeneity in the uptake of the intervention may increase the variance of the outcome in the intervention arm.

When we expect different ICCs in the two treatment arms, the variance of the treatment effect estimate is, assuming balanced data,

$$Var(\hat{\gamma}_1) = \frac{2\sigma_y^2}{n_1 n_2}\left\{[1 + (n_1 - 1)\rho_1] + [1 + (n_1 - 1)\rho_2]\right\}, \tag{1.32}$$

which simplifies to equation (1.11) when $\rho_1 = \rho_2$. This expression can be used in the formula for the noncentrality parameter to calculate power. Using the normal approximation, the total number of clusters required (with $n_2/2$ in each condition) can be calculated as

$$n_2 \geq 2\frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{n_1\delta^2}\left\{[1 + (n_1 - 1)\rho_1] + [1 + (n_1 - 1)\rho_2]\right\}. \tag{1.33}$$

1.4.1.5 Unequal allocation

Thus far, we have considered only trials with equal numbers of clusters in each arm. Unequal allocation can also be considered. Reasons that investigators may choose to use unequal allocation include reducing overall costs when one condition is more expensive to implement than the other, or to make trial participation more attractive by having a higher probability of being assigned to the intervention condition.

Let r denote the proportion of clusters allocated to arm 1. Then the numbers of clusters allocated to arms 1 and 2 are rn_2 and $(1-r)n_2$, respectively, and the variance of the treatment effect estimate becomes

$$Var(\hat{\gamma}_1) = \frac{\sigma_y^2\left[1 + (n_1 - 1)\rho\right]}{n_1 n_2}\left(\frac{1}{r} + \frac{1}{1-r}\right) = \frac{\sigma_\epsilon^2 + n_1\sigma_u^2}{n_1 n_2}\left(\frac{1}{r} + \frac{1}{1-r}\right). \tag{1.34}$$

The noncentrality parameter can be computed by using the square root of this expression for the denominator in equation (1.23) or (1.25).

Optimal allocation. For a given effect size, power will be maximized when $Var(\hat{\gamma}_1)$ is minimized. If we minimize equation (1.34) with respect to r, we obtain $r = 0.5$, which corresponds to equal allocation. However, the formula assumes that σ_y^2 and ρ, or σ_u^2 and σ_ϵ^2, are the same across arms. If we allow different ICCs and unequal allocation, the variance of the treatment effect estimate is

$$Var(\hat{\gamma}_1) = \frac{\sigma_y^2}{n_1 n_2}\left[\frac{1+(n_1-1)\rho_1}{r} + \frac{1+(n_1-1)\rho_2}{1-r}\right]. \tag{1.35}$$

The optimal allocation, optimal in the sense of minimizing the variance, can be found to be $\sqrt{d_1}/(\sqrt{d_1}+\sqrt{d_2})$, where the d_k are the design effects, $d_k = 1+(n_1-1)\rho_k, k = 1, 2$. For example, with ICCs of 0.02 and 0.06 for control and intervention and a cluster size of 50, the variance (and standard error) are minimized and the power is maximized if we allocate 41% of the clusters to the control condition and 59% to the intervention condition. In general, optimal allocation will involve allocating more clusters to the condition with the higher design effect (and thus higher variance).

Another optimal design strategy is to maximize cost efficiency, defined as the precision (inverse variance) of the treatment effect estimate divided by total study cost. Discussion of such designs can be found in [65]. Additional discussion of cluster randomized designs involving costs can be found in [41].

Unequal allocation of clusters to conditions

When the ICCs are equal in the two treatment arms, maximal power for a cluster randomized trial with a continuous outcome can be achieved by allocating clusters equally to each treatment arm. When the ICCs are different in the two arms, an optimal unequal allocation that maximizes power can be found. The optimal allocation will involve allocating more clusters to the condition with the higher design effect.

1.4.1.6 Covariates

Regression adjustment for covariates is often used in randomized trials to improve precision. In a linear regression model for independent observations, the effect of covariate adjustment is

$$\sigma_\epsilon^2 = \sigma_y^2(1 - R_{Y|X}^2) \tag{1.36}$$

where σ_y^2 is the total variance of Y, σ_ϵ^2 is the residual variance, and $R_{Y|X}^2$ is the proportion of the variance of Y that is explained by the covariates, represented by X. Because the residual variance is a key component of the standard errors of the regression coefficients, covariates that are strongly associated with the outcome variable can increase precision by reducing the residual variance.

Covariates might include demographic characteristics such as age and sex or clinical factors associated with prognosis such as cancer stage. One particularly notable covariate is the outcome variable measured at baseline; for example, a trial might administer a symptoms scale to patients at baseline and at follow up. In such trials, the data are typically analyzed using an analysis of covariance (ANCOVA), which tests for a mean difference between conditions in the outcome variable controlling for the baseline value of the variable.

Covariate adjustment can also be used in cluster randomized trials, but its effects are more complicated because of the multilevel data structure. In particular, the effects of covariates at the individual level and cluster level are different. We discuss covariates at each level.

Cluster-level covariates. A cluster-level covariate might be the gender or specialty of a health care provider in a trial randomizing providers, or the percentage of pupils below the poverty line in a trial randomizing schools. When a cluster-level covariate is added, the model becomes

$$Y_{ij} = \tilde{\gamma}_0 + \tilde{\gamma}_1 w_j + \tilde{\gamma}_2 z_j + \tilde{u}_j + \tilde{\epsilon}_{ij} \tag{1.37}$$

where tildes are used to indicate that the value of the regression coefficients and random terms may be different in the adjusted model. We denote the variance components in the adjusted model as $\tilde{\sigma}_u^2$ and $\tilde{\sigma}_\epsilon^2$. How do $\tilde{\sigma}_u^2$ and $\tilde{\sigma}_\epsilon^2$ compare to σ_u^2 and σ_ϵ^2? Since a cluster-level covariate takes the same value for all members of a cluster, it cannot explain variation among individuals within a cluster. Therefore the within-cluster variance is unchanged and $\tilde{\sigma}_\epsilon^2 = \sigma_\epsilon^2$. For the cluster-level variance, we have

$$\tilde{\sigma}_u^2 = (1 - \rho_B^2)\sigma_u^2, \tag{1.38}$$

where ρ_B is the between-cluster residual correlation between the outcome and the covariate. Using equation (1.38) and the relationships $\sigma_u^2 = \rho\sigma_y^2$, $\sigma_\epsilon^2 = (1-\rho)\sigma_y^2$ and $\sigma_y^2 = \sigma_u^2 + \sigma_\epsilon^2$, we can get an expression for the standard error of the treatment effect estimator as

$$SE(\hat{\tilde{\gamma}}_1) = \sqrt{\frac{4(\sigma_\epsilon^2 + n_1\tilde{\sigma}_u^2)}{n_1 n_2}} = \sqrt{\frac{4\sigma_y^2\left\{1 + [n_1(1-\rho_B^2) - 1]\rho\right\}}{n_1 n_2}}, \tag{1.39}$$

where the quantity in braces is the design effect when a cluster-level covariate is included, which simplifies to the standard design effect when $\rho_B = 0$. The total number of clusters required to achieve power of $1 - \beta$ is

$$n_2 \geq 4\frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{n_1\delta^2}\left\{1 + [n_1(1-\rho_B^2) - 1]\rho\right\}. \tag{1.40}$$

Because $1 - \rho_B^2 \leq 1$, we have $1 + [n_1(1-\rho_B^2) - 1]\rho \leq 1 + (n_1 - 1)\rho$; i.e., the design effect is reduced. Thus adjusting for a cluster-level covariate can be beneficial in reducing the required sample size. For example, for $\rho_B = 0.6$, $n_1 = 500$ and $\rho = 0.05$, the design effect is reduced from 25.95 to 16.95 by

including the covariate, a reduction of 35%. When $n_1 = 5$ or $n_1 = 50$, the reductions are 7.5% and 26%, respectively,

Including a cluster-level covariate comes at the expense of reducing the df in the t statistic for the test of the intervention effect. For each covariate term, one additional df is lost (recall that the df are $n_2 - q - 1$, where q is the number of cluster-level covariates, including the covariate encoding intervention assignment). This effect is negligible if the number of clusters is large, but can be important if the number of clusters is limited.

Individual-level covariates. Now we consider adjusting for an individual-level baseline covariate, denoted c_{ij}. The covariate could be the baseline measurement of the outcome variable, in which case the analysis becomes an ANCOVA, or it could be other covariates associated with the outcome. In this case, the relationships between the adjusted and unadjusted variances are [40]

$$\tilde{\sigma}_\epsilon^2 = \left(1 - \frac{n_1}{n_1 - 1}\rho_W^2\right)\sigma_\epsilon^2 \text{ and } \tilde{\sigma}_u^2 = \left(1 - \rho_B^2 + \frac{1}{n_1 - 1}\rho_W^2\right)\sigma_u^2 \qquad (1.41)$$

where ρ_W and ρ_B are the within-cluster and between-cluster residual correlations between the outcome Y_{ij} and the covariate c_{ij}, respectively; see [40]. These results show that the addition of the covariate can decrease the within-cluster variance but could increase or decrease the between-cluster variance; an increase occurs when $\rho_B^2 < (1/(n_1 - 1)\rho_W^2)$. This counterintuitive result is discussed in [56, 57]. However, when $\sigma_\epsilon^2 > \sigma_u^2$, any increase in σ_u^2 will be outweighed by the decrease in σ_ϵ^2 and overall the variance will be reduced [40]. This condition is expected to hold true for most CRTs.

The total required number of clusters can be found by using as the design effect

$$1 + [n_1(1 - \rho_B^2) - 1]\rho - \frac{n_1\rho_W^2(1 - 2\rho)}{n_1 - 1}. \qquad (1.42)$$

When individual-level covariates are entered into the model, no df are lost from the test of the intervention effect.

As an example, suppose the outcome variable is a symptoms score and the planned outcome analysis is an ANCOVA that adjusts for the baseline value of the score. The correlation between the baseline and follow up scores is expected to be about $\rho_W = 0.7$. If you lack a good estimate for the residual correlation between clusters (ρ_B), a conservative estimate is 0. For $n_1 = 10$ and $\rho = 0.05$, inclusion of the covariate would reduce the design effect from 1.45 to 1.045, corresponding to a reduction of 28% in the required sample size.

Somewhat different formulas for the effect of individual-level covariates have been presented by other authors [6, 48, 60].

Covariate adjustment in cluster randomized trials

Covariate adjustment can be an effective strategy to increase power in a cluster randomized trial. Both cluster-level and individual-level covariates that are correlated with the outcome variable can reduce the residual variance of the outcome and thereby increase the precision of the treatment effect estimator. Adding cluster-level covariates reduces the degrees of freedom for the test of the intervention effect. If the df available for the test of the intervention effect are limited (e.g., <30), cluster-level covariates should be restricted to those that are highly prognostic of the outcome to avoid loss of power.

1.4.1.7 Varying cluster sizes

Until now, we have assumed that all clusters have the same number of members. In practice, the sizes of clusters (schools, hospitals, villages, patients with the same health care provider) are likely to vary naturally. In addition, clusters that were of equal size at the beginning of a trial may experience non-response or dropout that leads to unequal cluster sizes in the final data set.

It has been shown that, given the same total number of clusters and number of participants, unequal cluster sizes are less efficient for estimating treatment effects than are equal cluster sizes [2]. Thus when cluster sizes vary, efficiency and power are reduced, and to achieve the same power, the sample size must be enlarged.

Let θ and τ denote the mean and standard deviation of the distribution of cluster sizes, respectively. An approximation of the relative efficiency of unequal versus equal cluster sizes is

$$RE \approx 1 - \lambda(1-\lambda)CV^2 \tag{1.43}$$

where $\lambda = \frac{\theta}{\theta+(1-\rho)/\rho}$ and CV is the coefficient of variation of the cluster size distribution, $CV = \tau/\theta$ [62]. The required total number of clusters can be obtained as the sample size assuming equal cluster sizes multiplied by $1/RE$ (note that the design effect is the inverse of the relative efficiency, and $1/RE$ is the design effect here). Other authors [1, 32, 36] have derived somewhat different estimates of the relative efficiency, all of which also depend on the ICC and CV of the cluster size distribution.

Investigators typically have a good idea of θ, the expected mean cluster size. To estimate the standard deviation of cluster size, a strategy is to estimate the minimum and maximum cluster size and approximate the standard deviation as one fourth of the range, i.e., $\tau \approx (\max - \min)/4$.

As an example, suppose that the mean cluster size is 16 and the ICC is 0.04. Standard deviations of cluster size of 0 (equal sizes), 4, 8 and 16 correspond to CVs of cluster size distribution of 0, 0.25, 0.5 and 1, and estimated relative

efficiencies of 1, 0.985, 0.94 and 0.76, respectively. The inverse REs are 1, 1.015, 1.064 and 1.32, meaning the sample sizes need to be inflated by 0%, 1.5%, 6.4% and 32%.

Unequal cluster sizes

When cluster sizes vary, efficiency and power are reduced. The loss of efficiency increases as the dispersion of cluster sizes, as measured by the coefficient of variation of the cluster size distribution, increases. The sample size required to achieve the desired level of power can be found by inflating the sample size requirement calculated assuming equal cluster sizes by the inverse of the relative efficiency, which can be approximated using equation (1.43).

1.4.1.8 Matching and stratification

In Section 1.2.1, we discussed how matching and stratification can be helpful in promoting balance between arms on prognostic factors. Stratification or matching prior to randomization can also improve power when these designs are used in conjunction with a stratified or matched analysis. In such an analysis, comparisons between conditions are made within strata. If clusters within strata are very similar, these comparisons will be akin to comparing the same experimental units under two different conditions. This reduces the between-cluster variability in the estimation of the intervention effect, reducing the standard error and increasing power.

The main impact of stratification in a CRT is to reduce the between-cluster variance component, σ_u^2 [15]. Because clusters in different conditions are now compared within strata, σ_u^2 is replaced with the variance among clusters within strata, which we denote σ_{um}^2. The variance of the treatment effect estimator becomes

$$Var(\hat{\gamma}_1) = \frac{4\left(\sigma_\epsilon^2 + n_1\sigma_{um}^2\right)}{n_1 n_2} \tag{1.44}$$

and the total number of clusters required is

$$n_2 \geq \frac{4(z_{1-\alpha/2} - z_{1-\beta})^2}{n_1\gamma_1^2}(\sigma_\epsilon^2 + n_1\sigma_{um}^2); \tag{1.45}$$

see [12]. Alternatively, we can define ρ_m as the correlation between cluster-level means within strata or matched pairs, equal to $\sigma_{um}^2/(\sigma_{um}^2 + \sigma_\epsilon^2)$, and the formula becomes

$$n_2 \geq \frac{4\sigma_y^2(z_{1-\alpha/2} - z_{1-\beta})^2}{n_1\gamma_1^2}[1 + (n_1 - 1)\rho - n_1\rho_m\rho]. \tag{1.46}$$

Thus the design effect is reduced by a factor of $n_1\rho_m\rho$. These formulae apply to studies with matched pairs and with larger strata.

Degrees of freedom will be lost when the stratification factors are used as covariates in a multilevel analysis model, which is needed in order to estimate the treatment effect within strata. Some authors suggest adding two additional clusters per study arm in matched-pair designs and 1-2 additional clusters per arm in designs with larger strata to account for the loss of df [23]. Because of the loss of df, for CRTs with a small number of clusters, pair matching should only be done if the clusters can be matched on factors that are highly correlated with the outcome variable; see [23] for further discussion, including a table showing the break-even values of the matching correlation above which pair matching will provide greater power than an unmatched trial.

Example. Suppose we wish to detect a small effect size of $\delta = 0.2$ with power of 80% and α of 0.05 two-sided, where clusters are of size $n_1 = 25$ and $\rho = 0.05$. We found in Section 1.4 that for an unstratified design, the design effect was 2.2 and the total number of clusters needed was 70. Suppose we consider stratification on cluster type and anticipate a correlation of $\rho_m = 0.3$ within strata. In this case, the design effect is reduced by $25 \times 0.3 \times 0.05 = 0.375$ and becomes 1.825. We will need a total of 1,433 individuals, or 58 clusters. Thus the number of clusters is reduced by 17%.

The matching correlation ρ_M can be difficult to predict. In the face of uncertainty about ρ_M, a conservative approach is to ignore any gain in power due to stratification.

Effect of matching and stratification on power

In a stratified design, comparisons between conditions are made within strata. If the strata are homogeneous, the between-cluster variance relevant to estimating the treatment effect is reduced. Thus matching and stratification can increase power and reduce sample size requirements, in addition to promoting balance on baseline covariates. However, this variance reduction comes with a loss of df for estimating the treatment effect.

1.4.2 Dichotomous outcomes

In this section, sample size and power formulas for cluster randomized trials with dichotomous outcomes are derived using logic similar to that for CRTs with continuous outcomes. In the section on analysis, we discussed two hierarchical models for binary data, the cluster-level proportions model and the cluster-level log odds model. For sample size and power, we focus on the cluster-level proportions model, which has parameters that are easier to understand. Since many concepts were previously discussed for continuous outcomes, our discussion of dichotomous outcomes is more brief.

1.4.2.1 Sample size and power

Using the cluster-level proportions model, the intervention effect can be estimated as the difference in sample proportions, $\hat{\pi}_1 - \hat{\pi}_2$, and its variance, also given in (1.16), is

$$Var(\hat{\pi}_1 - \hat{\pi}_2) = \left[\frac{\pi_1(1-\pi_1)}{n_1 n_2/2} + \frac{\pi_2(1-\pi_2)}{n_1 n_2/2}\right][1 + (n_1 - 1)\rho_d]\,. \qquad (1.47)$$

where ρ_d is as defined in (1.14). For large samples, the test statistic $(\hat{\pi}_1 - \hat{\pi}_2)/\sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)}$ has a standard normal distribution under the null hypothesis. Using this approach, the total number of clusters required to achieve power of $1-\beta$ with two-sided α of 0.05 when cluster size is n_1 is

$$n_2 \geq \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)][1 + (n_1-1)\rho_d]}{n_1(\pi_1 - \pi_2)^2}, \qquad (1.48)$$

This formula assumes equal allocation, constant cluster sizes and equal ICCs in each arm. These assumptions are relaxed in later sections. This formula can also be derived by calculating the total sample size requirement across both arms, N, for independent observations,

$$N = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{(\pi_1 - \pi_2)^2}, \qquad (1.49)$$

and inflating by the design effect. Some authors recommend adding one extra cluster per treatment arm [23]. Power for a given cluster size and number of clusters can be obtained by solving the equation for $z_{1-\beta}$ and applying the standard normal cumulative distribution function. R functions for sample size and power computation are given below.

```
sampsize.crt.bal.bin <- function(p1, p2, rho, n1, beta, alpha){
  za <- qnorm(1-alpha/2)
  zb <- qnorm(1-beta)
  num <- 2*(za+zb)^2*(p1*(1-p1)+p2*(1-p2))*(1+(n1-1)*rho)
  denom <- n1*(p1-p2)^2
  n2 <- num/denom
  print("Total number of clusters required is")
  print(n2)
  print("Required clusters per arm is")
  print(ceiling(n2/2))
}
power.crt.bal.bin <- function(p1, p2, rho, n1, n2, alpha){
  za <- qnorm(1-alpha/2)
  num <- n1*n2*(p1-p2)^2
  denom <- 2*(p1*(1-p1)+p2*(1-p2))*(1+(n1-1)*rho)
  zb <- sqrt(num/denom) - za
  pnorm(zb)
}
```

Example. Suppose that we anticipate proportions of 0.3 and 0.5 in the two arms and clusters each have 25 members. The ICC is estimated to be 0.03. Using the command

```
sampsize.crt.bal.bin(.3, .5, .03, 25, 0.2, 0.05)
```

the trial is estimated to need 7 clusters per condition to achieve at least 80% power with two-sided α of 0.05. The actual power can be computed using the command

```
power.crt.bal.bin(.3, .5, .03, 25, 14, 0.05)
```

which indicates that the actual power is 84.5%.

1.4.2.2 Sample size per cluster

Given a fixed total number of clusters n_2, the required cluster size is

$$n_1 = \frac{(1-\rho_d)[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)](z_{1-\alpha/2}+z_{1-\beta})^2}{(\pi_1-\pi_2)^2 n_2/2 - \rho_d[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}. \tag{1.50}$$

As mentioned for continuous outcomes, it will not always be possible to achieve desired power with a fixed number of clusters, even if the sample size per cluster is extremely large.

1.4.2.3 Unequal ICCs in treatment arms

In some cases we may wish to use separate ICC estimates in the two arms. For example, we may anticipate more dispersion of cluster-level proportions in the intervention arm, or we may want to account for the fact that the ICC depends on the underlying proportion as $Var(\hat{p})/[\pi(1-\pi)]$; see (1.14). Separate ICCs by arm can be incorporated into the sample size formula to yield

$$n_2 \geq \frac{2(z_{1-\alpha/2}+z_{1-\beta})^2[\pi_1(1-\pi_1)Deff_1]+\pi_2(1-\pi_2)Deff_2]}{n_1(\pi_1-\pi_2)^2}, \tag{1.51}$$

where $Deff_k = 1+(n_1-1)\rho_{dk}$ is the design effect in arm k.

1.4.2.4 Unequal allocation

If the number of clusters allocated to arms 1 and 2 are rn_2 and $(1-r)n_2$, respectively, the variance of the risk difference becomes

$$Var(\hat{\pi}_1-\hat{\pi}_2) = \frac{\pi_1(1-\pi_1)d_1}{rn_1n_2} + \frac{\pi_2(1-\pi_2)d_2}{(1-r)n_1n_2} \tag{1.52}$$

where the d_k are the design effects, $d_k = 1+(n_1-1)\rho_k, k = 1,2$; we have allowed unequal ICCs and thus unequal design effects for generality. As discussed for continuous outcomes, the optimal allocation that minimizes the variance of the treatment effect estimator is $\sqrt{d_1}/(\sqrt{d_1}+\sqrt{d_2})$.

1.4.2.5 Covariates

The impact of cluster-level and individual-level covariates on the power of CRTs with continuous outcomes was discussed in Section 1.4.1.6. The impact of adjusting for covariates in trials with binary outcomes is more complex. Due to the nonlinearity of the logistic regression model, it is difficult to derive tractable expressions for the variances in adjusted models [55]. Furthermore, in a logistic regression model, the unadjusted and adjusted treatment effect parameters differ. Unadjusted analyses yield marginal estimates that compare an intervention subject with a randomly selected control subject. Adjusted analyses yield conditional estimates that compare an intervention subject to a control subject with the same covariate values. For continuous outcomes, the adjusted and unadjusted treatment effects are the same, but this is not generally true for binary outcomes [22]. In a logistic model, the adjustment typically increases the estimated treatment effect; that is, estimated odds ratios will be further from 1 (where an odds ratio of 1 indicates no treatment effect). Furthermore, including covariates in a logit model tends to *increase* the variance of the estimated treatment effect in log odds terms [52].

For these reasons, simple formulas that account for the impact of covariates on power in a CRT with a binary outcome analyzed using a mixed effects logit model are lacking. However, Schochet [55] provides an approach that is based on using a GEE estimator rather than the mixed effects logit model. A key finding is that gains in power due to covariate adjustment are likely to be smaller for binary outcomes than they are for continuous outcomes.

1.4.2.6 Varying cluster sizes

Varying cluster sizes are less efficient for estimating treatment effects than are equal cluster sizes, as previously discussed. The same approach for accounting for this reduction in efficiency can be used for both continuous and binary outcomes; see Section 1.4.1.7 for more details.

1.5 Additional resources

Books on the design and analysis of cluster randomized trials include Murray [43], Donner and Klar [15], Eldridge and Kerry [17], Campbell and Walters [8] and Hayes and Moulton [23]. Some of these books, e.g. [23], discuss time-to-event outcomes, rates and counts. Survival outcomes are also discussed in [31]. Power analysis for trials with multilevel data, including cluster randomized trials, multicenter trials and individually randomized group treatment trials, is discussed in Moerbeek and Teerenstra [41]. Sample size calculation for clustered and longitudinal outcomes are discussed in Ahn et

al. [1]. Some recent reviews summarize key results for sample size calculations for CRTs [18, 53].

Many journals require that reports of trials conform to the guidelines in the Consolidated Standards of Reporting Trials (CONSORT) statement. There is a CONSORT statement extension specifically for cluster randomized trials [9] that provides a checklist of items to include in the trial report, including the ICC, which researchers often neglect to report [13].

1.5.1 Resources for other designs

This chapter has covered outcome analysis and sample size and power for some common designs of cluster randomized trials. We discuss several major designs and provide references.

Individually randomized group treatment trials. In an individually randomized group treatment trial, individuals are randomized to study conditions but receive their intervention with other participants, typically in a group setting, or through a change agent shared with other participants. For example, in a mindful awareness intervention for breast cancer survivors, participants randomized to the intervention were assigned to groups who attended classes together [7]. In these studies, there is little or no group-level ICC at baseline, but a positive ICC is expected among the outcomes of individuals within the same group. Special methods are needed for analysis and sample size estimation for these studies. Literature discussing these studies includes [4, 46, 45, 41, 51].

Cluster-randomized crossover trials. In a simple crossover trial, each subject receives each treatment in random order. Because each subject serves as his or her own control, a crossover design can be quite powerful. In a cluster randomized crossover trial, clusters are randomly allocated to a sequence of interventions. Two designs can be distinguished: crossover at the cluster level, in which each subject is included in only one of the treatment periods, and crossover at the subject level, in which each subject is observed in both periods [50]. Crossover CRTs are discussed in [20, 50, 49], and a brief summary of sample size formulas is provided in [12].

Stepped wedge trials. A stepped wedge trial is similar to a crossover trial except that the crossovers are all in one direction, from control to intervention condition, and are staggered over time. Clusters are randomized to cross over to the intervention at time points called steps, and all clusters end the trial in the intervention condition. References for stepped wedge trials include [3, 24, 25, 64].

1.5.2 Resources for power and sample size calculation

The National Institutes of Health has a website with guidance on research methods related to studies that randomize groups or clusters or that deliver in-

terventions to groups at https://researchmethodsresources.nih.gov. The website includes a sample size calculator.

The free software program Optimal Design Plus Empirical Evidence includes power and sample size calculation for more complex design elements such as three or four levels. The program and documentation are available at http://www.wtgrantfoundation.org.

Moerbeek and Teerenstra [41] describe the SPA-ML (Statistical Power Analysis for Multi-Level designs) program, which is available for free download at http://tinyurl.com/SPAML. Their book describes the use of the program, which currently only handles continuous outcomes.

Campbell and Walters [8] discuss both data analysis and power and sample size for CRTs, and provide code in R, Stata and SPSS. The code is available at their website http://sheffield.ac.uk/scharr/sections/dts/statistics.

Bibliography

[1] C. Ahn, M. Heo, and S. Zhang. *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. CRC Press/Taylor and Francis Group, Boca Raton, FL, 2015.

[2] B.E. Ankenman, A.I. Aviles, and J.C. Pinheiro. Optimal designs for mixed-effects models with two random nested factors. *Statistica Sinica*, 13:385–401, 2003.

[3] G. Baio, A. Copas, G. Ambler, J. Hargreaves, E. Beard, and R.Z. Omar. Sample size calculation for a stepped wedge trial. *Trials*, 16(1):354, 2015.

[4] S.A. Baldwin, D.J. Bauer, E. Stice, and P. Rohde. Evaluating models for partially clustered designs. *Psychological Methods*, 16(2):149–65, 2011.

[5] R. Bastani, B. A. Glenn, A.E. Maxwell, A.M. Jo, A.K. Herrmann, C. M. Crespi, W.K. Wong, L.C. Chang, S.L. Stewart, T.T. Nguyen, M.S. Chen, and V.M. Taylor. Cluster-Randomized Trial to Increase Hepatitis B Testing among Koreans in Los Angeles. *Cancer Epidemiology, Biomarkers and Prevention*, 24(9):1341–9, 2015.

[6] H.S. Bloom, L. Richburg-Hayes, and A.R. Black. Using covariates to improve precision for studies that randomized schools to evaluate educational interventions. *Evaluation and Policy Analysis*, 29:30–59, 2007.

[7] J.E. Bower, A.D. Crosswell, A.L. Stanton, C. M. Crespi, D. Winston, J. Arevalo, J. Ma, S.W. Cole, and P.A. Ganz. Mindfulness meditation for younger breast cancer survivors: a randomized controlled trial. *Cancer*, 121(8):1231–40, 2015.

[8] M.J. Campbell and S.J. Walters. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health-Related Research*. John Wiley & Sons Ltd, Chichester, UK, 2014.

[9] M.K. Campbell, G. Piaggio, D.R. Elbourne, D.G. Altman, and C. Group. CONSORT 2010 statement: extension to cluster randomised trials. *British Medical Journal*, 345:e5661, 2012.

[10] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 2nd edition, 1988.

[11] J.A. Cook, T. Bruckner, G.S. MacLennan, and C.M. Seiler. Clustering in surgical trialsdatabase of intracluster correlations. *Trials*, 13:2, 2012.

[12] C.M. Crespi. Improved designs for cluster randomized trials. *Annual Review of Public Health*, 37:1–16, 2016.

[13] C.M. Crespi, A.E. Maxwell, and S. Wu. Cluster randomized trials of cancer screening interventions: are appropriate statistical methods being used? *Contemporary Clinical Trials*, 32:477–84, 2011.

[14] P.J. Diggle, P.K. Heagerty, K.Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, Oxford, 2nd edition, 2002.

[15] A. Donner and N. Klar. *Design and Analysis of Cluster Randomization Trials in Health Research*. Oxford University Press, New York, New York, USA, 2000.

[16] S.M. Eldridge, O.C. Ukoumunne, and J.B. Carlin. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. *International Statistical Review*, 77:378–94, 2009.

[17] Kerry S. Eldridge S. *A Practical Guide to Cluster Randomized Trials in Health Research*. Arnold, London, 2012.

[18] F. Gao, A. Earnest, D.B. Matchar, M.J. Campbell, and D. Machin. Sample size calculations for the design of cluster randomized trials: a summary of methodology. *Contemporary Clinical Trials*, 42:41–50, 2015.

[19] J.C. Gardiner, Z.H. Luo, and L.A. Roman. Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, 28:221–39, 2009.

[20] B. Giraudeau, P. Ravaud, and A. Donner. Sample size calculation for cluster randomized corss-over trials. *Statistics in Medicine*, 27:5578–85, 2008.

[21] E.M. Hade, D.M. Murray, M.L. Pennell, D. Rhoda, E.D. Paskett, V.L. Champion, B.F. Crabtree, A. Dietrich, M.B. Dignan, M. Farmer, J.J. Fenton, S. Flocke, R.A. Hiatt, S.V. Hudson, M. Mitchell, P. Monahan, S. Shariff-Marco, S.L. Slone, K. Stange, S.L. Stewart, and P.A.O. Strickland. Intraclass Correlation Estimates for Cancer Screening Outcomes: Estimates and Applications in the Design of Group-Randomized Cancer Screening Studies. *JNCI Monographs*, 2010:97–103, 2010.

[22] W.W. Hauck, S. Anderson, and S.M. Marcus. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials*, 19:249–56, 1998.

[23] R.J. Hayes and L.H. Moulton. *Cluster Randomised Trials*. Taylor & Francis Group, LLC, Boca Raton, FL, USA, 2nd edition, 2017.

[24] K. Hemming, T.P. Haines, P.J. Chilton, A.J. Girling, and R.J. Lilford. The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *BMJ*, 350:h391, 2015.

[25] K. Hemming and M. Taljaard. Sample size calculations for stepped wedge and cluster randomised trials: A unified approach. *Journal of Clinical Epidemiology*, 69:137–46, 2016.

[26] J.J. Hox. *Multilevel Analysis: Techniques and Applications*. Routledge, New York, New York, USA, 2nd edition, 2009.

[27] F.B. Hu, J. Goldberg, D. Hedeker, B.R. Flay, and M.A. Pentz. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, 147:694–703, 1998.

[28] N.M. Ivers, I.J. Halperin, J. Barnsley, J.M. Gromshaw, B.R. Shah, K. Tu, R. Upshur, and M. Zwarenstein. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials*, 13:120, 2012.

[29] N.M. Ivers, M. Taljaard, S. Dixon, C. Bennett, A. McRae, J. Taleban, Z. Skea, J.C. Brehaut, R.F. Boruch, M.P. Eccles, J.M. Grimshaw, C. Weijer, M. Zwarenstein, and A. Donner. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 20008. *BMJ*, 343:343, 2011.

[30] P. Jahn, O. Kuss, H. Schmidt, A. Bauer, M. Kitzmantel, K. Jordan, S. Krasemann, and M. Landenberger. Improvement of pain-related self-management for cancer patients through a modular transitional nursing intervention: A cluster-randomized multicenter trial. *Pain*, 155:746–54, 2014.

[31] A. Jahn-Eimermacher, K. Ingel, and A. Schneider. Sample size in cluster-randomized trials with time to event as the primary endpoint. *Statistics in Medicine*, 32:739–51, 2013.

[32] S.M. Kerry and J.M. Bland. Unequal cluster sizes for trials in English and Welsh general practices: implications for sample size calculations. *Statistics in Medicine*, 20:377–90, 2001.

[33] Y. Kim, Y.-K. Choi, and S. Emery. Logistic Regression with Multiple Random Effects: A Simulation Study of Estimation Methods and Statistical Packages. *American Statistician*, 67:171–82, 2013.

[34] L. Kish. *Survey Sampling*. John Wiley, New York, New York, USA, 1965.

[35] K.Y. Liang and S.L. Zeger. LongitudinL data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

[36] A.K. Manatunga, M.G. Hudgens, and S. Chen. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal*, 43:75–86, 2001.

[37] O. Manor and D.M. Zucker. Small sample inference for the fixed effects in the mixed linear model. *Computational Statistics and Data Analysis*, 46:801–17, 2004.

[38] A.E. Maxwell, L.L. Danao, R.T. Cayetano, C.M. Crespi, and R. Bastani. Implementation of an evidence-based intervention to promote colorectal cancer screening in community organizations: A cluster randomized trial. *Translational Behavioral Medicine*, 6:295–305, 2016.

[39] S.I. Mishra, R. Bastani, C.M. Crespi, L.C. Chang, P.H. Luce, and C.R. Baquet. Results of a randomized trial to increase mammogram usage among Samoan women. *Cancer Epidemiology, Biomarkers and Prevention*, 16(12):2594–604, 2007.

[40] M. Moerbeek. Power and money in cluster randomized trials: When is it worth measuring a covariate? *Statistics in Medicine*, 25:2607–17, 2006.

[41] M. Moerbeek and Teerenstra S. *Power Analysis of Trials with Multilevel Data*. Taylor & Francis Group, LLC, Boca Raton, FL, USA, 2016.

[42] M. Moerbeek and M.P.F. Van Breukelen, G.J.P. amd Berger. Optimal experimental designs for multilevel logistic models. *The Statistician*, 50(1):1–14, 2001.

[43] D.M. Murray. *Design and Analysis of Group-Randomized Trials*. Oxford University Press, New York, NY, USA, 1998.

[44] D.M. Murray and J.L. Blitstein. Methods to reduce the impact of intraclasscorrelation in group-randomized trials. *Evaluation Review*, 27(1):79–103, 2003.

[45] S.P. Pals, D.M. Murray, C.M. Alfano, W.R. Shadish, P.J. Hannan, and W.L. Baker. Erratum. *American Journal of Public Health*, 98(12):2120, 2008.

[46] S.P. Pals, D.M. Murray, C.M. Alfano, W.R. Shadish, P.J. Hannan, and W.L. Baker. Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. *American Journal of Public Health*, 98(8):1418–24, 2008.

[47] G.M. Raab and I. Butcher. Balance in cluster randomized trials. *Statistics in Medicine*, 20(3):351–65, 2001.

[48] S.W. Raudenbush. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2:173–85.

[49] N.G. Reich, J.A. Myers, D. Obeng, A.M. Milstone, and T.M. Perl. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLoS One*, 7:e35564, 2012.

[50] C. Rietbergen and M. Moerbeek. The design of cluster randomized crossover trials. *Journal of Educational and Behavioral Statistics*, 36(4):472–90, 2011.

[51] C. Roberts and S.A. Roberts. Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, 2:152–62.

[52] L.D. Robinson and N.P. Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Reviews*, 58:227–240, 1991.

[53] C. Rutterford, A. Copes, and S. Eldridge. Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 44(3):1051–67, 2015.

[54] R. Sankaranarayanan, B.M. Nene, S.S. Shastri, K. Jayant, R. Muwonge, A.M. Budukh, S. Hingmire, S.G. Malvi, R. Thorat, A. Kothari, R. Chinoy, R. Kelkar, S. Kane, S. Desai, V.R. Keskar, R. Rajeshwarkar, N. Panse, and K.A. Dinshaw. HPV screening for cervical cancer in rural India. *New England Journal of Medicine*, 360(14):1385–94, 2009.

[55] P.Z. Schochet. Statistical power for school-based RCTs with binary outcomes. *Journal of Research on Educational Effectiveness*, 6:263–94, 2013.

[56] T.A.B. Snijders and R.J. Bosker. Modeled variance in two-level models. *Sociological Methods and Research*, 22(3):342–63, 1994.

[57] T.A.B. Snijders and R.J. Bosker. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications Ltd. London, England, UK, second edition, 2012.

[58] J. Spybrook, H. Bloom, R. Congdon, C. Hill, A. Martinez, and S. Raudenbush. *Optimal Design Plus Empirical Evidence: Documentation for the Optimal Design Software*, 2011.

[59] D.R. Taves. Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics*, 15:443–53.

[60] S. Teerenstra, S. Eldridge, M. Graff, E. de Hoop, and G.F. Borm. A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine*, 31:2169–78.

[61] R.M. Turner, R.Z. Omar, and S.G. Thompson. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine*, 20:453–72.

[62] G.J.P. van Breukelen, M.J.J.M. Candel, and M.P.F. Berger. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*, 26:2589–603, 2007.

[63] E.H. Wagner, E.J. Ludman, E.J. Aiello Bowles, R. Penfold, R.J. Reid, C.M. Rutter, J. Chubak, and R. McCorkle. Nurse navigators in early cancer care: a randomized, controlled trial. *Journal of Clinical Oncology*, 32(1):12–8, 2014.

[64] W. Woertman, E. de Hoop, M. Moerbeek, S.U. Zuidema, D.L. Gerritsen, and S. Teerenstra. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology*, 66(7):752–8, 2013.

[65] S. Wu, W.K. Wong, and C.M. Crespi. Maximin optimal designs for cluster randomized trials. *Biometrics*, 73(3):916–26, 2017.